Adaptive Open-Set Active Learning with Distance-Based Out-of-Distribution Detection for Robust Task-Oriented Dialog System

Sai Keerthana Goruganthu
University of Missouri
sgmhz@umsystem.edu

Roland Oruche
University of Missouri
rro2q2@umsystem.edu

Prasad Calyam
University of Missouri
calyamp@missouri.edu

Abstract

The advancements in time-efficient data collection techniques such as active learning (AL) have become salient for user intent classification performance in task-oriented dialog systems (TODS). In realistic settings, however, traditional AL techniques often fail to efficiently select targeted in-distribution (IND) data when encountering newly acquired out-of-distribution (OOD) user intents in the unlabeled pool. In this paper, we introduce a novel adaptive open-set AL framework viz., "AOSAL" for TODS that combines a distance-based OOD detector using an adaptive false positive rate threshold along with an informativeness measure (e.g., entropy) to strategically select informative IND data points in the unlabeled pool. Specifically, we utilize the adaptive OOD detector to classify and filter out OOD samples from the unlabeled pool, then prioritize the acquisition of classified IND instances based on their informativeness scores. To validate our approach, we conduct experiments that display our framework's flexibility and performance over multiple distance-based approaches and informativeness measures against deep AL baselines on benchmark text datasets. The results show that our AOSAL consistently outperforms the baselines on IND classification and percentage of acquired IND samples, demonstrating its ability to improve robustness of task-oriented dialog systems.

1 Introduction

Recent advances in time-efficient data collection techniques such active learning (AL) (Settles, 2009; Ren et al., 2021) show the promise of significantly improving the performance of task-oriented dialog system (TODS) for tasks related to user intent classification (Zhang and Zhang, 2019; Wu et al., 2024). The time-efficient AL techniques not only improve the model accuracy of the TODS, but also help reduce the annotation budget of human annotators when querying the most informative samples that accelerate training performance.

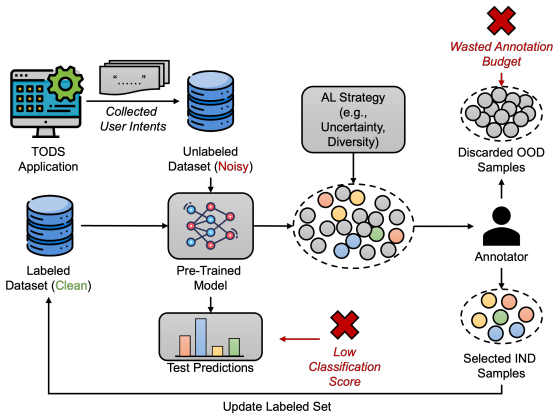


Figure 1: The challenges of traditional AL methods when encountering OOD instances from newly collected user intents in the unlabeled pool which includes low classification score and wasted annotation budget.

In real-world applications, however, existing AL methods often struggle to select in-distribution (IND) data from unlabeled pools containing out-of-distribution (OOD) user intents, leading to inefficiencies in the learning process. Figure 1 illustrates the challenge of employing standard AL frameworks in a TODS application, where an unlabeled dataset of collected user intents are noisy due to instances that are OOD. Typical queries using e.g., uncertainty (Lewis, 1995) and diversity-based methods (Nguyen and Smeulders, 2004) are prone to selecting a high number of OOD instances, which in turn can waste the annotation budget of the human annotator. Consequently, this can lead to low classification performance, and more concretely, incorrect dialog responses if there are insufficient amount of IND samples selected for training, as shown in the example scenario in Figure 2.

Previous works have investigated robust AL frameworks in the context of open-set recognition (Scheirer et al., 2012). The work in (Yang et al., 2024) develops a progressive active learn-

Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 357–369
September 18–20, 2024.

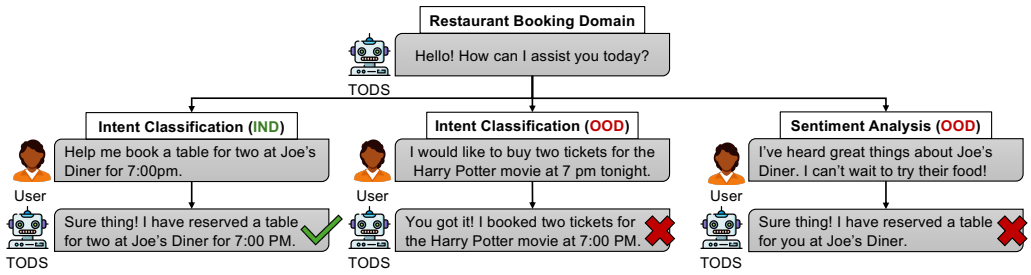


Figure 2: Example scenarios of task-oriented dialog systems (TODS) handling user intents in a restaurant booking domain. TODS can provide incorrect and unwarranted responses when encountering OOD intents.

ing framework that implements an OOD detector for filtering OOD instances in the unlabeled pool. Although other works have proposed similar methods related to open-set active learning (Du et al., 2021; Park et al., 2022; Ning et al., 2022), these works are mainly applied in the computer vision space, and are unsuited for NLP tasks in TODS. In addition, while AL frameworks in the NLP space such as CAL (Margatina et al., 2021) and CounterAL (Deng et al., 2023) address OOD generalization challenges, they are not practical to open-set AL where the unlabeled dataset contains a mixture of IND and OOD samples. Given the emergence and applicability of TODS in various application domains e.g., healthcare, banking, there presents a need to develop robust AL strategies that avoid OOD instances while also acquiring informative IND instances that improve model training.

In this paper, we present a novel adaptive open-set AL framework viz., "AOSAL" for TODS that combines an adaptive distance-based OOD detector with informative sampling measures (e.g., uncertainty, diversity) to effectively acquire IND samples in the unlabeled pool. Our OOD detector features a normalized score function that classifies unlabeled samples based on their distance to each class in the IND labeled space. We enable our OOD detector to be sensitive to distribution shifts by employing an adaptive threshold that is controlled by using a predetermined false positive rate (FPR) over the OOD detection performance. Based on the prioritization of classified pseudo-IND samples, we then leverage sampling measures for selecting the most informative instances for annotation. In addition, we demonstrate the flexibility of our AOSAL approach to multiple distance-based functions (Podolskiy et al., 2021; Frogner et al., 2015) and informative measures (Lewis, 1995).

We perform experiments to validate our AOSAL framework over four NLP benchmark related to intent classification (Larson et al., 2019; Gangal et al., 2020), and sentiment analysis (Maas et al., 2011; Aslam et al., 2020), comparing its performance against several deep AL baselines that are based on uncertainty, diversity, and hybrid-based approaches. Experimental results suggest that our AOSAL approach consistently outperforms the baselines on metrics such IND classification accuracy, and percentage of acquired IND/OOD samples.

The remainder of the paper is organized as follows: Section 2 describes related work. In Section 3, we detail the AOSAL methodology. In Section 4, we detail the experimental setup and provide the main results against AL baselines. An analysis to test the robustness of AOSAL is presented in Section 5. Section 6 discusses the limitations of our approach, and finally, Section 7 concludes our work.

2 Related Work

2.1 Active Learning

Recent advancements in active learning have leveraged pool-based sampling (Settles, 2009), where an agent can select and query a large set of instances to the oracle (i.e., human annotator) from the unlabeled pool. Common methods on the selection process, or *query strategy*, based on how informative a given sample is, include uncertainty (Lewis, 1995; Settles, 2009) and diversity (Nguyen and Smeulders, 2004; Sener and Savarese, 2018) methods. Uncertainty strategies such as Entropy (Settles, 2009) and Least Confidence (Lewis, 1995) aim to select a set of instances from the unlabeled pool in cases where the model is least confident in its prediction. While uncertainty can maintain low computational complexity, diversity-based methods such as Coreset (Sener and Savarese, 2018) and clustering-based methods (Nguyen and Smeulders, 2004) select samples that better represent the distribution of the unlabeled pool.

The advent of deep learning in AL has en-

abled batch-mode active learning (Kirsch et al., 2019), where the sampling of unlabeled instances in batches are sent to the oracle for labeling. Authors in (Kirsch et al., 2019) extend Bayesian Active Learning by Disagreement (BALD) (Houlsby et al., 2011), by presenting BatchBALD, which alleviates the time complexity of calculating the mutual information between an individual sample and the model parameters. Batch-model AL has also engendered recent work in hybrid-based approaches (Yin et al., 2017; Zhdanov, 2019; Ash et al., 2020; Shui et al., 2020). The work in (Ash et al., 2020) develops BADGE, a hybrid query strategy that robustly selects samples by leveraging both the prediction uncertainty and diverse samples from the hallucinated gradient space of the model. Despite such advancements, these methods often fail to improve IND classification performance and limit the oracle's annotation budget when there is a distribution mismatch between the labeled and the unlabeled set. Thus, traditional active learning methods are not feasible for training agents within TODS systems used in real-world applications.

2.2 Open-Set Active Learning

Previous works have aimed to develop AL methods in the context of open-set recognition (Scheirer et al., 2012) that is more suitable for realistic scenarios where there presents a distribution mismatch in the unlabeled pool (Kothawade et al., 2021; Du et al., 2021; Ning et al., 2022; Park et al., 2022; Safaei et al., 2024; Yang et al., 2024). The work in (Du et al., 2021) develops CCAL, which utilizes contrastive learning to extract semantic and distinctive features in the unlabeled pool. The authors propose an AL error when selecting invalid (OOD) samples, which are segmented between valid and invalid query errors. Other works such as in (Kothawade et al., 2021) develop a unified AL framework that addresses OOD samples in the unlabeled pool by utilizing submodular conditional mutual information that jointly models the similarity between the query set and batch of unlabeled samples and their dissimilarity between a conditioning set.

More recent work on open-set AL further addresses distribution mismatches by utilizing OOD samples for training in the unlabeled pool. For instance, progressive active learning (PAL) (Yang et al., 2024) samples both pseudo-IND and pseudo-OOD samples to simultaneously train the ID classifier and a proposed OOD detector using a one-vs-all classifier. Authors in (Park et al., 2022) demonstrate that balancing between purity (i.e., distinguishing between collected IND and OOD instances), and informativeness (i.e., uncertainty, diversity) consistently improves the classifier accuracy under various noise (OOD) ratio in the unlabeled pool. Similarly, other works such as LfOSA (Ning et al., 2022) and EOAL (Safaei et al., 2024) leverage both known (IND) and unknown (OOD) data instances to effectively informative IND samples while avoid OOD samples during AL rounds. Despite these advancements, the majority of methods from existing work in open-set AL are mainly tailored to the computer vision domain.

Our AOSAL framework for robust TODS is novel because it: (i) detects OOD instances (e.g., user intents) using distance-based approaches coupled with an adaptive threshold to maintain a low false positive rate in text-based datasets, and (ii) utilizes measures over unlabeled instances classified as IND for improving IND accuracy on the labeled set. In addition, we demonstrate that our AL framework can be extended to multiple distance-based approaches and informative measures.

3 Methodology

In this section, we describe the problem formulation for open-set AL and then detail the overview and components of our proposed AOSAL framework.

3.1 Problem Formulation

We define a TODS problem for identifying user intents as a $\mathcal{K}$-class classification task. An IND labeled dataset $\mathcal{D}_L$ has an input space $\mathcal{X}$ and a corresponding output label space $\mathcal{Y} \in \{1, \ldots, \mathcal{K}\}$ of $\mathcal{K}$ IND classes, which are independently and identically distributed (i.i.d.) from $\mathcal{D}_L$. The full dataset is defined as $\mathcal{D}_L = \{(x_i, y_i)\}_{i=1}^{N_L}$, where N_L is the length of the initial labeled training set.

We denote an unlabeled dataset as $\mathcal{D}_U$ for the purposes of re-training the TODS over newly collected user intents. Formally, the unlabeled dataset is defined as $\mathcal{D}_U = \{(x_j)\}_{j=1}^{N_U}$, where N_U is the length of the unlabeled set. We also denote $N_L << N_U$, highlighting the substantially larger pool of unlabeled dataset $\mathcal{D}_U$ compared to $\mathcal{D}_L$. In real-world AL scenarios, there often presents a distribution mismatch in the unlabeled pool due noisy, OOD class samples. Thus, we define our problem to an open-set AL in a pool-based setting, where

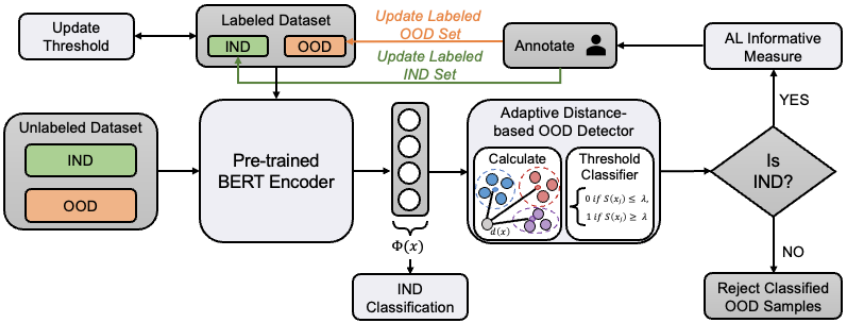


Figure 3: Main Architecture of our AOSAL framework. A pre-trained BERT model encodes samples unlabeled from the unlabeled pool and classifies them using the OOD detector. Classified IND samples are queried using an informative measure and fed to the annotator for updating the labeled set and updating the FPR-controlled threshold.

the unlabeled pool contains a both IND and OOD input samples (i.e., $\mathcal{D}_U = \mathcal{X}^{IND} \cup \mathcal{X}^{OOD}$) with a label space $\mathcal{Y}^{IND}$ and $\mathcal{Y}^{OOD}$, respectively, and $\mathcal{Y}^{IND} \cap \mathcal{Y}^{OOD} = \emptyset$. In simple terms, a sample $x_j \in \mathcal{D}_U$ may belong to an IND or OOD class in the unlabeled dataset.

Within the AL loop, an AL strategy queries a batch of samples of size b from $\mathcal{D}_U$ to form into a query set Q, which can consist a mixture of IND and OOD samples (i.e., $Q = \mathcal{D}_U^{IND} \cup \mathcal{D}_U^{OOD}$). This query set is then fed to the human annotator (i.e., oracle) for labeling and updating the initial training set $\mathcal{D}_L$.

3.2 Adaptive Open-Set Active Learning

We present our novel adaptive open-set AL (AOSAL) framework that couples an adaptive distance-based OOD detector with informativeness measures for efficiently managing OOD instances in the unlabeled pool. We display the main architecture of our AOSAL approach in Figure 3. Unlike previous AL frameworks for NLP (Margatina et al., 2021; Deng et al., 2023), we utilize the unlabeled OOD instances that are queried to the oracle for annotation to improve our distance-based OOD detector with an adaptive threshold controlled by a pre-defined false positive rate (FPR). In the following, we formalize the main components of AOSAL and detail the full sampling procedure in the AOSAL cycle.

3.2.1 Distance-based OOD Detector

To address the challenge of efficiently utilizing annotation resources in AL contexts, we have developed a distance-based OOD detector. This detector classifies an unlabeled sample x_j as either in-distribution (IND) or out-of-distribution (OOD) based on an adaptive threshold. The classification decision is made according to the following rule:

$$\text{Classify}(x_j) = \begin{cases} \textbf{accept}, & \text{if } S(x_j) \leq \lambda, \\ \textbf{reject}, & \text{if } S(x_j) > \lambda \end{cases} \quad (1)$$

where, λ is the threshold in the range [0, 1] and x_j denotes the j-th unlabeled sample in $\mathcal{D}_U$. The threshold separates IND samples, which score at or below the threshold, from OOD samples, which score above it. The scoring function $S(x_j)$ is designed to measure the proximity of x_j to the nearest class in the labeled dataset $\mathcal{D}_L$.

The scoring function $S(x_j)$ is conceptualized to enhance the selection of IND samples from the unlabeled dataset $\mathcal{D}_U$ by calculating the minimal distances between x_j and each class represented in $\mathcal{D}_L$. It is defined as:

$$S(x_j) = \min_{k \in \mathcal{K}} d(x_j, \mu_{x_k}), \quad (2)$$

where, μ_{x_k} represents the mean feature vector of class k from the set of classes $\mathcal{K}$ in $\mathcal{D}_L$. This approach ensures that the scoring function remains adaptable across various distance metrics, each potentially having different mathematical properties and score ranges.

To facilitate uniformity in classification regardless of the absolute scale of distance values, we normalize the scores to a [0, 1] range:

$$S(x_j) = \frac{S(x_j)}{\max_{x_j \in \mathcal{D}_U}(S(x_j))}. \quad (3)$$

This normalization not only standardizes the score across various distance metrics but also aligns with the thresholding approach to identify between IND and OOD samples. In our experiments, we

utilize the Mahalanobis distance (Podolskiy et al., 2021) and Wasserstein distance (Frogner et al., 2015) to compute $S(x_j)$, which are chosen for their robustness in capturing the geometric nuances of data distributions. The specific formulas and their application are detailed further in the Appendix A.1, ensuring a comprehensive exposition of our distance-based OOD detection methodology.

3.2.2 Adaptive Threshold

Classifying OOD instances using a constant threshold value creates significant challenges in maintaining high OOD accuracy in real-world settings. This is particularly evident in newly collected unlabeled data in TODS applications, where IND and potential OOD samples can cause a distribution shift. Consequently, this can lead to high false positives (i.e., detecting OOD samples as IND) and ultimately negatively impact the annotation budget as more OOD samples are naturally acquired.

To address this, we implement an adaptive threshold mechanism controlled by a pre-defined false positive rate (FPR), which is essential for maintaining classification integrity under varying data conditions. The FPR is defined as the ratio of IND instances mistakenly classified as OOD to the total number of true negative instances. It is calculated as:

$$FPR = \frac{FP}{FP + TN} \tag{4}$$

where, FP is the number of false positives, (i.e., IND samples incorrectly classified as OOD), and TN is the number of true negatives, IND (i.e., samples correctly classified as IND).

To maintain system accuracy and adapt to new data, the adaptive threshold λ is adjusted based on the FPR, which is calculated as:

$$\lambda = FPR(\mathcal{D}_{\text{val}}, \alpha) \tag{5}$$

where, $\mathcal{D}_{\text{val}}$ the validation set with a mixture of labeled IND and OOD samples (i.e., $\mathcal{D}_{\text{val}} = \mathcal{D}_L^{IND} \cup \mathcal{D}_L^{OOD}$), and α is the predetermined FPR rate. The benefit of ensuring an adaptive threshold is consistent with the pre-defined FPR that mitigates the risk of the OOD detector from producing high false positives on the unlabeled dataset during AL acquisition. Furthermore, λ is dynamically calibrated to ensure that the proportion of false positives does not exceed α. This dynamic adjustment is conducted through a meticulous analysis of the model's scoring outputs on each validation sample $x_{\text{val}} \in \mathcal{D}_{\text{val}}$. The threshold λ is then set such that:

$$\alpha = \frac{|\{x_{val} \in \mathcal{D}_L : S(x_{\text{val}}) > \lambda \text{ and } y_{\text{val}} = 0\}|}{|\{x_{\text{val}} \in \mathcal{D}_L : y_{\text{val}} = 0\}|} \tag{6}$$

where, $S(x_{\text{val}})$ is the score function applied to each validation sample, and y_{val} indicates the sample's label, with a label of 0 signifying an IND sample.

The validation set plays a crucial role in accurately updating the adaptive threshold for effective OOD detection in TODS. The informativeness metric derived from calculating uncertainty or diversity on the validation set is utilized to fine-tune the model and threshold. Furthermore, the validation set is continuously updated with newly annotated OOD samples, ensuring that the OOD detector remains up-to-date and capable of handling evolving data patterns. This mechanism enhances the robustness and reliability of TODS, enabling them to maintain high accuracy in OOD detection under varying situations and adapt to dynamic data shifts.

3.3 AOSAL Sampling Procedure

The overall AL sampling process for our proposed AOSAL framework is shown in Algorithm 1. We start by training a deep learning model M_θ on $\mathcal{D}_L$ at the initial iteration $t = 0$ to obtain M_{θ_t}. During the validation, we leverage our distance-based OOD detector that computes a score from Equation 3 over samples on the validation set. The normalized scores from the OOD detector are then used to set the initial threshold λ_t based on a pre-defined FPR α over the validation set $\mathcal{D}_{\text{val}}$.

Within our AL loop, we extract the features for each unlabeled sample in $\mathcal{D}_U$ computed by M_{θ_t} as input to our OOD detector using a normalized distance-based function from Equation 3 that computes the distance based scores for classification. After classifying the samples based on Equation 1, we ignore the classified OOD samples and focus on acquiring IND samples using informative measures (e.g., uncertainty, diversity). Following this, human annotators refine these classifications, and the resulting samples consisting of both IND and potential OOD are updated in either the IND train set or the OOD validation set.

The iteration of the model is updated at $t = t+1$, and the threshold is adjusted using the OOD detector with a controlled false positive rate (FPR) at α. This process is repeated until the annotation budget

Algorithm 1 Adaptive Open-Set Active Learning with Distance-Based OOD Detection

Require: Labeled IND dataset D_L^{IND}, labeled OOD dataset D_L^{OOD}, unlabeled pool D_U, validation set D_{val}, model M_θ, encoder function Φ, acquisition size b, labeling budget B, total query set Q, threshold function FPR, informativeness measure U, current iteration t.

Train $M_{\theta_{t=0}}$ on $\mathcal{D}_L^{IND}$ for multi-classification
$\lambda_{t=0} \leftarrow FPR(D_{\text{val}}, \alpha)$ ▷ Set initial threshold (Eq. 5)
while $|Q| < B$ **do**
  for each x_j in D_U **do**
    $\mu_{x_k} \leftarrow \Phi(x_k), k = \{1, ..., \mathcal{K}\}$
    $s_{x_j} \leftarrow S(\Phi(x_j), \mu_{x_k})$ ▷ From Eq. 3
    if $s_{x_j} \leq \lambda$ **then** ▷ IND label
      $A \leftarrow \{(x_j^{IND}, \hat{y}_j^{IND})\}$
    end if
  end for
  for each x_j in A **do**
    $Q \leftarrow \underset{x_j \in U}{\text{argmax}} U(x_j), |Q| = b$ ▷ Select b instances with highest informative scores.
  end for
  $D_L^{IND} \leftarrow D_L^{IND} \cup \{Q^{IND} \setminus D_U\}$ ▷ Update train set with acquired IND samples
  $D_{\text{val}}^{\text{OOD}} \leftarrow D_{\text{val}}^{\text{OOD}} \cup \{Q^{OOD} \setminus D_U\}$ ▷ Update validation set with acquired OOD samples
  Train the model $M_{\theta_{t+1}}$ on D_L^{IND}
  Update λ using Eq. 5 on updated D_{val}
  $t \leftarrow t + 1$
end while
return M_{θ_t}, λ_t ▷ Return updated model and threshold

B is exhausted, ensuring continuous refinement of the model's performance and the threshold.

4 Experiments and Results

In this section, we provide our experimental setup for open-set AL on benchmark NLP datasets and provide the main results of our AOSAL approach against baseline AL datasets.

4.1 Datasets

We validated our AOSAL framework over NLP datasets related to topic classification and sentiment analysis. These datasets are integral for validating the model's efficacy in handling both IND and OOD samples within varied textual contexts. For topic classification, we test over the CLINC-Full (Larson et al., 2019) dataset with 150 classes and Real Out-of-distribution Sentences From Task-Oriented Dialog (ROSTD) (Gangal et al., 2020) dataset with 12 classes, which both include OOD samples. For sentiment analysis, we utilize the Stanford Sentiment Treebank (SST)-2 (Aslam et al., 2020) dataset with only 2 classes each for the positive and negative sentiments. In our experiments, we set one dataset to the IND class and the other dataset to the OOD class. For instances where CLINC-Full and ROSTD are assigned to the IND class, we join the remaining OOD samples along with the assigned OOD dataset. We provide the full dataset description and partitions in Appendix A.2.

4.2 Baselines and Implementation Details

We compare our approach against five AL baselines that include state-of-the-art approaches for different query strategies. Specifically, we test an uncertainty sampling method, namely Entropy (Joshi et al., 2009), for which samples with the lowest confidence in the model's predictive probability are selected. For diversity sampling, we test our approach against BERT – KM from the works of (Yuan et al., 2020), where they performs k-means clustering over the L2-normalized BERT embeddings to select diverse samples in the unlabeled feature space. For hybrid sampling, we compare our approach against BADGE (Ash et al., 2020), which is known to be an AL state-of-the-art method. In addition to state-of-the-art AL methods, we include CAL (Margatina et al., 2021) that selects "contrastive" unlabeled samples based on their feature similarity and divergent predictive probability. Lastly, we include Random sampling as a baseline for randomly acquiring instances in the unlabeled pool.

We implement our approach using a pre-trained BERT model (Devlin et al., 2019) from the HuggingFace library [1] as the backbone model for each approach in our experiments. While we opted to use BERT due to its reliable performance on natural language understanding tasks, our AOSAL framework can be extended to multiple model architectures for intent classification (Liu et al., 2019; Lan et al., 2019; He et al., 2021).

For each dataset, we use 10% of the train set as our initial labeled set $\mathcal{D}_L$ and use 10% of OOD samples and label them in the validation set for OOD detection. In addition, we set the noise ratio (i.e.,

[1] HuggingFace BERT model available at: https://huggingface.co/bert-base-uncased

percentage of OOD samples in the unlabeled pool) to 30%. This noise ratio presents a realistic consideration of the amount of noise that can be present in the unlabeled pool. During each AL iteration, we fine-tune $\mathcal{D}_L$ with newly acquired IND samples from the unlabeled pool $\mathcal{D}_U$. We set the oracle labeling budget B to 25% percent of $\mathcal{D}_U$ for a total number of 5 AL rounds. We pre-trained the base model over 5 epochs on CLINC-Full for training. We run experiments for each AL method 5 times over each dataset and report the average IND accuracy and the percentage of IND/OOD samples in the acquisition size $|Q|$ for each AL iteration. We provide additional detail on the model implementation using BERT and relevant hyperparameters in Appendix A.3.

4.3 Main Results

The main results over the CLINC-Full (IND) and the SST-2 (OOD) dataset are shown in Figure 4. Figures 4a and 4b show the average test accuracy (90.1% ($\pm$0.01)) and the average acquired IND (661.38 ($\pm$6.66)) of our AOSAL approach across all AL iterations compared to the baselines, respectively. When averaging across all AL budgets, our approach shows significant improvement in acquiring IND samples compared to Entropy (57.82 ($\pm$6.11)). The relatively low test accuracy performance for uncertainty methods such as Entropy (87.3% ($\pm$0.004)) may be the result of selecting instances in the unlabeled pool that are least confident in its prediction, which causes Entropy to acquire more OOD instances, thus wasting the annotation budget.

In the scope of diversity- and hybrid-based methods, AOSAL shows comparable test accuracy performance to BADGE (90% ($\pm$0.00)) and BERT-KM (91.1% ($\pm$0.01)). While our AOSAL approach significantly shows higher acquired IND compared to BADGE (388.40 ($\pm$3.94)) and BERT-KM (462.96 ($\pm$8.62)), the high test accuracy results may indicate the benefit of acquiring a diverse set of samples in the unlabeled pool for improving model performance. In addition, CAL shows surprisingly low performance in both IND test accuracy (87.8% ($\pm$0.01)) and average acquired IND (76.56 ($\pm$10.65)) when handling OOD instances from the SST-2 dataset.

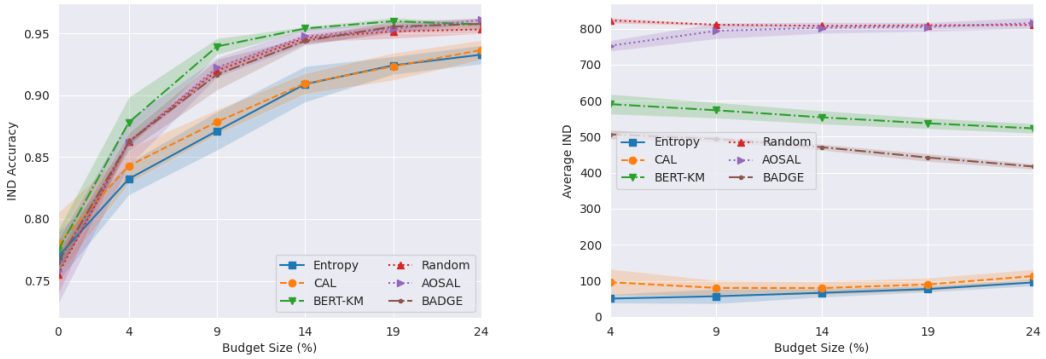
Furthermore, AOSAL shows comparable results in acquired IND to Random sampling (676.71 ($\pm$2.32)) and an improvement in IND test accuracy performance (89.8% ($\pm$0.01)). Since Random AL sampling follows a uniform distribution, it outperforms all baseline approaches when the amount of OOD instances in the unlabeled pool is considerably low (i.e., noise ratio at 30%). Despite this, the acquired IND performance does not always translate to high IND test accuracy, as indicated in Figure 4a (89.8% ($\pm$0.01)). This is because the samples acquired may not always be informative in terms of uncertainty and diversity for effectively improving model performance.

Similar results on the consistency of AOSAL are shown in Figure 5. AOSAL shows comparable average performance to the baselines across all AL iterations in terms of IND test accuracy (91.2% ($\pm$0.01)) in Figure 5a and acquired IND (681.20 ($\pm$4.90)) in Figure 5b. Compared to the baselines such as Entropy (470.13 ($\pm$61.09)), Random (678.52 ($\pm$2.28)), and CAL (426.217 ($\pm$35.91)), our AOSAL approach shows a higher amount of acquired IND averaged across all AL iterations. This in turn translates to comparable or higher accuracy on the IND test set. While the IND test accuracy results are comparable to other baselines such as BADGE (91.8% ($\pm$0.01)) and BERT-KM (91.8% ($\pm$0.00)), our AOSAL approach maintains a comparable accuracy as well as average acquired IND performance to BADGE (662.683 ($\pm$10.31)) and BERT-KM (702.122 ($\pm$4.98)) when encountering a variety of OOD instances in the unlabeled pool.

5 Analysis

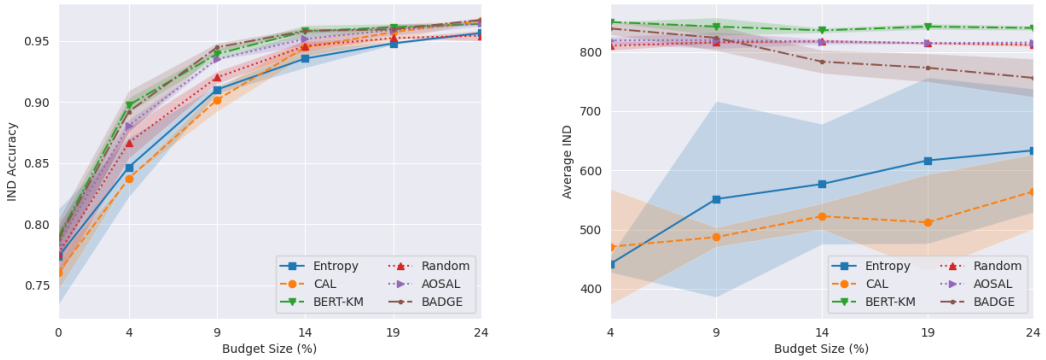
5.1 Ablation Study

We conduct an ablation study to check the AOSAL framework under varying budgets, analyzing how different distance metrics and OOD detection can influence IND accuracy. We compare six configurations of our framework, including AOSAL-CONST, which uses a constant threshold, and others such as AOSAL-NO-OOD, AOSAL-MAH-DIV, AOSAL-WAS-DIV, AOSAL-MAH-UNC, and AOSAL-WAS-UNC that use an adaptive FPR threshold, but differ in their application of Mahalanobis or Wasserstein distances and the incorporation of uncertainty and diversity metrics. To ensure fair and meaningful comparisons across all experimental settings, we utilize CLINC-Full as the IND data and SST-2 as the OOD data, with a fixed false positive rate (FPR) of 95%. This standardization helps maintain consistent experimental conditions throughout the study. Figure 6 shows that the con-



(a) CLINC (IND) and SST-2 (OOD) on IND test accuracy. (b) CLINC (IND) and SST-2 (OOD) on acquired IND.

Figure 4: Test accuracy results and averaged acquired IND on the each AL method over the CLINC-Full (IND) and SST-2 (OOD) dataset. Each method was ran 5 times with different seeds and the average accuracies were reported.



(a) CLINC (IND) and ROSTD (OOD) IND test accuracy. (b) CLINC (IND) and ROSTD (OOD) on acquired IND.

Figure 5: Test accuracy results and averaged acquired IND on the each AL method over the CLINC-Full (IND) and ROSTD (OOD) dataset. Each method was ran 5 times with different seeds and the average accuracies were reported.

figurations lacking OOD detection i.e., AOSAL-NO-OOD demonstrates a significant drop in the model's performance, highlighting the crucial role of effective OOD detection mechanisms in enhancing the overall accuracy of the system. This analysis confirms the robustness and versatility of our AOSAL framework in adapting to different operational constraints and validates the utility of advanced distance measures for OOD detection in the AL environment.

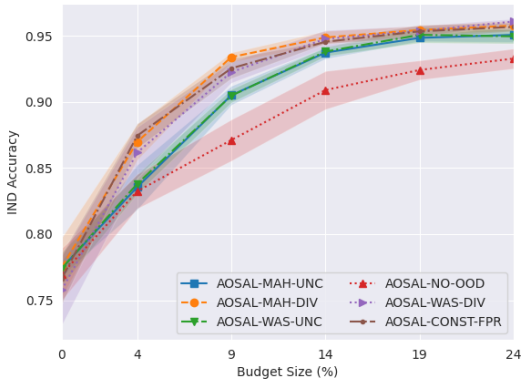


Figure 6: Ablation study on the IND test accuracy over the CLINC-Full (IND) and SST-2 (OOD) dataset using different AOSAL variants.

5.2 Threshold Analysis

We conducted a detailed threshold analysis to evaluate the impact of various FPR thresholds on IND accuracy. Our study systematically explored the performance implications of different FPR levels including 90%, 95%, and 97% across multiple datasets. The dataset configurations, detailed in Table 1, included CLINC-Full as the IND dataset paired with ROSTD and SST-2 as OOD datasets. These combinations were selected to rigorously evaluate the robustness of our AOSAL approach across diverse scenarios. The results clearly indicate that the IND accuracy is sensitive to the FPR threshold set. For the CLINC-Full and ROSTD dataset configuration, the IND accuracy peaks at a

FPR (%)	IND ACC (%)	
	CLINC (IND) ROSTD (OOD)	CLINC (IND) SST-2 (OOD)
90	95.71 (±0.00)	95.28 (±0.01)
95	95.93 (±0.01)	**96.14** (±0.01)
97	**96.15** (±0.00)	95.31 (±0.00)

Table 1: IND accuracy at different FPR thresholds for CLINC-Full (IND), ROSTD (OOD), and SST-2 (OOD).

97% FPR setting, suggesting a balanced threshold that avoids excessive false positives while maintaining a high detection rate of in-domain samples. Conversely, tightening the FPR to 95% shows a slight dip in accuracy, which could imply an over-restriction misses some IND instances. A similar trend is observed in the CLINC-Full and SST-2 dataset configuration, reinforcing the importance of carefully calibrating the FPR threshold according to specific dataset characteristics and operational requirements. This analysis underscores the significance of the AOSAL's adaptability to different operational scenarios. By systematically evaluating various FPR thresholds, we can identify the optimal setting that balances the trade-off between maintaining high in-domain accuracy and minimizing false positives.

6 Limitations

Sensitivity to Hyperparameters. One of the key challenges of our AOSAL framework is its dependence on hyperparameter settings. The choice of hyperparameters such as adaptive threshold for FPR and informative parameters is critical for achieving maximal learning efficiency. However, reaching this balance is by nature difficult since this directly affects the framework's performance in correctly distinguishing OOD samples. Getting the wrong values for hyperparameters leads to either underconfidence or overconfidence in OOD instances and hence the model's overall performance. Future works can be directed towards implementing more intelligent adaptive hyper-parameter tuning methods that are sensitive to changes in the data environment.

Model Performance with Sparse Data. Another critical limitation arises when there is a lack of data availability. With few input data points, our framework cannot generate and calibrate the right distance metrics for OOD detection. This can hinder the accurate classification and enhancement of OOD detection, especially in the initial stages of training the model. There are potential ways to tackle these challenges, such as improving data augmentation methods and the use of synthetic data generation to help improve the model's performance despite starting with minimal initial data.

Scalability in Human-in-the-loop Setting. While oracles enable AI models of TODS to train more efficiently with fewer samples via annotations, this process is not always scalable for annotators. This challenge in a human-in-the-loop setting is particularly evident when oracles provide a significant number of annotations for OOD samples within each AL round due to large unlabeled pools. Alternatively, previous works have created modeling approaches in other domains such as computer vision (Ning et al., 2022; Yang et al., 2024; Safaei et al., 2024) that train over both IND/OOD samples and AL sampling techniques for automatically extracting OOD samples in the unlabeled pool. Consequently, this effectively reduces the number of annotations the oracle provides.

7 Conclusion

In this paper, we presented AOSAL which is an AL framework that aims to improve the efficiency and effectiveness of TODS. AOSAL combines a distance-based OOD technique with an adaptive FPR threshold and an informativeness measure based on uncertainty and diversity. This integration enables AOSAL to improve the classification of IND and OOD samples and thus focuses primarily on the most useful IND examples from an unlabeled data pool for training. The experimental analysis we have conducted shows that AOSAL is highly effective for dealing with complex datasets in comparison to traditional active learning techniques. These real-world applications have demonstrated the practical usefulness and effectiveness of the framework in enhancing not only the robustness but also the accuracy of intent classification in TODS by the AOSAL framework's ability to selectively acquire high-value IND training samples.

In future work, one can investigate advanced data augmentation and synthetic data approaches to facilitate training in data-deficient scenarios and design adaptive hyperparameter optimization of the system's responsiveness to data variability.

References

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. In *Proceedings of ICLR*.

Andleeb Aslam, Usman Qamar, Pakizah Saqib, Reda Ayesha, and Aiman Qadeer. 2020. A novel framework for sentiment analysis using deep learning. In *2020 22nd International Conference on Advanced Communication Technology (ICACT)*, pages 525–529. IEEE.

Xun Deng, Wenjie Wang, Fuli Feng, Hanwang Zhang, Xiangnan He, and Yong Liao. 2023. Counterfactual active learning for out-of-distribution generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11362–11377.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Pan Du, Suyun Zhao, Hui Chen, Shuwen Chai, Hong Chen, and Cuiping Li. 2021. Contrastive coding for active learning under class distribution mismatch. In *Proceedings of IEEE/CVF*.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. 2015. Learning with a wasserstein loss. In *Advances in neural information processing systems*, volume 28.

Varun Gangal, Abhinav Arora, Arash Einolghozati, and Sonal Gupta. 2020. Likelihood ratios and generative classifiers for unsupervised out-of-domain detection in task oriented dialog. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7764–7771.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.

Kuan-Hao Huang. 2021. Deepal: Deep active learning in python. *arXiv preprint arXiv:2111.15258*.

Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. 2009. Multi-class active learning for image classification. In *2009 ieee conference on computer vision and pattern recognition*, pages 2372–2379. IEEE.

Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. 2019. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32.

Suraj Kothawade, Nathan Beck, Krishnateja Killamsetty, and Rishabh Iyer. 2021. Similar: Submodular information measures based active learning in realistic scenarios. *Advances in Neural Information Processing Systems*, 34:18685–18697.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of EMNLP-IJCNLP*.

David D Lewis. 1995. A sequential algorithm for training text classifiers: Corrigendum and additional data. In *Proceedings of ACM SIGIR*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. In *Proceedings of EMNLP*.

Hieu T Nguyen and Arnold Smeulders. 2004. Active learning using pre-clustering. In *Proceedings of ICML*.

Kun-Peng Ning, Xun Zhao, Yu Li, and Sheng-Jun Huang. 2022. Active learning for open-set annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 41–49.

Dongmin Park, Yooju Shin, Jihwan Bang, Youngjun Lee, Hwanjun Song, and Jae-Gil Lee. 2022. Meta-query-net: Resolving purity-informativeness dilemma in open-set active learning. *Advances in Neural Information Processing Systems*, 35:31416–31429.

Alexander Podolskiy, Dmitry Lipin, Andrey Bout, Ekaterina Artemova, and Irina Piontkovskaya. 2021. Revisiting mahalanobis distance for transformer-based out-of-domain detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13675–13682.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. 2021. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40.

Bardia Safaei, VS Vibashan, Celso M de Melo, and Vishal M Patel. 2024. Entropic open-set active learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4686–4694.

Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. 2012. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1757–1772.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

Burr Settles. 2009. Active learning literature survey. *University of Wisconsin-Madison Department of Computer Sciences*.

Changjian Shui, Fan Zhou, Christian Gagné, and Boyu Wang. 2020. Deep active learning: Unified and principled method for query and training. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pages 1308–1318. PMLR.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP*.

Yuxia Wu, Tianhao Dai, Zhedong Zheng, and Lizi Liao. 2024. Active discovering new slots for task-oriented conversation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Yang Yang, Yuxuan Zhang, Xin Song, and Yi Xu. 2024. Not all out-of-distribution data are harmful to open-set active learning. *Advances in Neural Information Processing Systems*, 36.

Changchang Yin, Buyue Qian, Shilei Cao, Xiaoyu Li, Jishang Wei, Qinghua Zheng, and Ian Davidson. 2017. Deep similarity-based batch mode active learning with exploration-exploitation. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 575–584. IEEE.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. In *Proceedings of EMNLP*.

Leihan Zhang and Le Zhang. 2019. An ensemble deep active learning method for intent classification. In *Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence*, pages 107–111.

Fedor Zhdanov. 2019. Diverse mini-batch active learning. *arXiv preprint arXiv:1901.05954*.

A Appendix

A.1 Generalization of Distance-based OOD Detection Method

As introduced in Section 3.2.1, the distance score function $S(x_j)$ is designed to be adaptable to various distance metrics, accommodating different mathematical properties and score ranges. Specifically, for a $\mathcal{K}$-class classification problem, we maximize the selection of IND samples from $\mathcal{D}_U$ by computing the minimum distance between an unlabeled sample x_j and each class in the labeled dataset $\mathcal{D}_L$. Herein, we demonstrate applicability of our generalized distance-based OOD detector to the Mahalanobis distance (Podolskiy et al., 2021) and Wasserstein distance (Frogner et al., 2015).

Mahalanobis Distance. We utilize the Mahalanobis distance has shown to be useful for classifying detecting OOD instances without the reliance accessing OOD instances for training (Podolskiy et al., 2021). This distance method is a way to determine the closeness of an data sample to a set of data samples that belongs to a class k.

Given an unlabeled sample x_j from $\mathcal{D}_U$, the Mahalanobis distance can be calculated as:

$$d(x_j) = \min_{k \in \mathcal{K}} (\Phi(x_j) - \mu_{x_k})^\top \Sigma^{-1} (\Phi(x_j) - \mu_{x_k}), \tag{7}$$

where $\Phi(x_j)$ is the embedding of the unlabeled sample x_j, μ_{x_k} is the mean of the multivariate Gaussian distribution of class $k \in \{1, ..., \mathcal{K}\}$, and Σ represents the covariance matrix. The calculations of μ_k and Σ are computed as:

$$\mu_{x_k} = \frac{1}{N_k} \sum_k \Phi(x), \tag{8}$$

$$\Sigma = \frac{1}{N_L} \sum_k \sum_{i \in k} (\Phi(x_i) - \mu_{x_k})(\Phi(x_i) - \mu_{x_k})^\top, \tag{9}$$

where N_k is the number of training samples the class k and N_L is the total number of training samples in the labeled set. While the range of distances of the Mahalanobis distance is [0, ∞], we transform the ranges Equation 7 to [0, 1] using Equation 3.

Wasserstein Distance. Similarly, the Wasserstein distance calculates the minimal cost of transporting mass from the distribution of x_j to that of each class distribution k where the cost is defined by the ground distance between the distributions (Frogner et al., 2015). Given an unlabeled sample x_j from $\mathcal{D}_U$, the Wasserstein distance can be calculated as:

$$S(x_j) = \arg\min_{k \in \mathcal{K}} W(\Phi(x_j), \mu_{x_k}) \tag{10}$$

$$W(\Phi(x_j), \mu_k) = \inf_{\gamma \in \Gamma(P_{\Phi(x_j)}, P_{\mu_{x_k}})} \int \|\Phi(x_j) - \mu_k\|_2 \, d\gamma(\Phi(x_j), \mu_{x_k}) \tag{11}$$

Here, $\Phi(x_j)$ is the feature vector of x_j, $P_{\Phi(x_j)}$ and P_{μ_k} represent the empirical distributions of x_j and class k, respectively, and $\Gamma(P_{\Phi(x_j)}, P_{\mu_k})$ contains all feasible joint distributions γ where the marginals are $P_{\phi(x_j)}$ and P_{μ_k}. Waterstein distances have a non-negative range [0, ∞], where 0 represents perfect similarity between distributions. These distances can be normalized to the range [0, 1], using a transformation similar to Equation 3.

A.2 Dataset Details

In this section, we provide the dataset statistics of each NLP benchmark dataset shown in Table 2. In the following, we provide a brief description for each of the dataset as it related to intent classification.

CLINC-Full. The CLINC-Full dataset was introduced by (Larson et al., 2019) which is designed for intent classification across multiple domains such as banking, home, travel, and business. It contains a total of 23,700 queries, out of which 22,500 are in-distribution (IND) queries spanning 150 classes for intent classification tasks, and 1,200 are out-of-distribution (OOD) samples for out-of-scope prediction. This dataset is crucial for assessing the model's capability to classify complex, real-world user intents and includes numerous OOD scenarios to evaluate robustness in model performance.

ROSTD. The Real Out-of-domain Sentences From Task-Oriented Dialog (ROSTD) dataset, proposed by (Gangal et al., 2020), is designed for training and evaluating intent classification models in task-oriented dialog systems with a focus on out-of-distribution robustness. It contains 34,059

Statistic	CLINC-Full	ROSTD	SST-2
Train	16950	25218	54577
Valid	2700	3537	6822
Test	4050	5304	6822
OOD samples	1200	4590	0
% of OOD samples in unlabeled pool	7.87%	20.22%	0%
IND classes	150	12	2

Table 2: Dataset statistics for CLINC-Full, ROSTD and SST-2.

queries across 12 classes, including in-distribution queries and an additional 4,590 out-of-distribution samples curated with human annotations. The dataset aims to facilitate the development of more robust dialog systems capable of handling out-of-distribution utterances effectively.

SST-2. The Stanford Sentiment Treebank (SST-2) (Aslam et al., 2020) is another well-established benchmark for sentiment analysis, particularly for tasks that involve considering sentence structure and sentiment polarity. It consists of 67,314 sentences for training, 855 for validation, and 1,821 for testing, all derived from movie review sentences on Rotten Tomatoes. Each sentence is labeled as positive, negative, or neutral.

A.3 Model Implementation & Hyperparameters

In this section, we provide details of the model implementation and hyperparameters used in our experiments. We use a pre-trained BERT model (Devlin et al., 2019) from the HuggingFace library (Wolf et al., 2020) and integrated it in our Python environment using PyTorch 2.0 and PyTorch Lightning. We train BERT using a batch size of 32, learning rate of $5e-5$, AdamW optimizer epsilon $1e-6$ and weight decay of 0.001, and embedding dimension of 768. For all datasets, we used a maximum sequence length of 256. We pre-trained the base model over 5 epochs on CLINC-Full and 1 epoch on ROSTD and SST-2. In the AL cycle, we use the newly acquired samples from $\mathcal{D}_U$ to fine-tune BERT over the updated labeled set $\mathcal{D}_L$. We ensure fair comparison among each AL method by evaluating them 5 times using a different random seed. Each experiment is run on an Nvidia A100 80GB GPU. We use the open source materials from (Huang, 2021; Ash et al., 2020; Margatina et al., 2021) to implement the baseline AL methods from their respective source code repository on GitHub.